ABSTRACT

An improved system for streaming a software application to a plurality of clients comprises a principal server having the software stored thereon as a plurality of blocks and a plurality of intermediate servers between the principal server and the clients. The principal server is configured to stream program and data blocks to downstream devices in accordance with a dynamic prediction of the needs of those devices. The intermediate servers are configured to cache blocks received from connected upstream devices and service requests for blocks issued from downstream devices. In addition, the intermediate servers are further configured to autonomously predict the needs of downstream devices, stream the predicted blocks to the downstream devices, and if the predicted blocks are not present in the intermediate server cache, request those blocks from upstream devices. The intermediate servers can also be configured to make intelligent cache purging decisions with reference to the contents of the caches in other connected devices.